Вып. 18. — С.27-30.

10. http://mathhelpplanet.com/static.php?p=onlain-reshit-treugolnik

FINDING SOME ELEMENTS OF THE TRIANGLE IN THE SCA MAPLE

A.M. Nigmedzianova

The procedure for finding the equations of the sides, medians, altitudes, the coordinates of the points of intersection of an arbitrary triangle in the SCM Maple.

Keywords: computer modelling, analytical geometry, solutions of triangles, equations of sides, medians, heights of a triangle.

УДК 530.12

ПРИМЕНЕНИЕ БРИТАНСКОГО КОРПУСА АКАДЕМИЧЕСКОГО ПИСЬМЕННОГО АНГЛИЙСКОГО ЯЗЫКА В КУРСЕ «ВВЕДЕНИЕ В ОБРАБОТКУ ЕСТЕСТВЕННОГО ЯЗЫКА» ДЛЯ СТУДЕНТОВ ИТ-СПЕЦИАЛЬНОСТЕЙ

А.Б. Нугуманова[1], Е.М. Байбурин[2]

1 *yalisha@yandex.kz*; Восточно-Казахстанский государственный университет им. С. Аманжолова, Усть-Каменогорск, Республика Казахстан
2 *ebaiburin@vkgu.kz*; Восточно-Казахстанский государственный университет им. С. Аманжолова, Усть-Каменогорск, Республика Казахстан

В данной работе описывается применение Британского корпуса академического письменного английского языка в качестве машиночитаемого ресурса, иллюстрирующего возможности современных методов обработки естественного языка. Британский корпус представляет собой хорошо сбалансированную и репрезентативную электронную коллекцию, составленную из высококачественных академических текстов по 35 учебным дисциплинам.

Ключевые слова: обработка естественного языка, информационный поиск, корпус.

Британский корпус академического письменного английского языка, сокращенно именуемый BAWE (The British Academic Written English), создавался как совместный проект трех британских вузов: Уорикского университета, Университета Рединга и Университета Оксфорд Брукс. Цель проекта состояла в том, чтобы собрать в единый корпус лучшие образцы письменных работ студентов-старшекурсников и магистрантов указанных вузов [1]. Таким образом, в корпус вошло около 3000 работ по 35 учебным дисциплинам, представляющим 4 области наук: искусство и гуманитарные науки, науки о жизни, физические науки и социальные науки. В настоящее время корпус доступен для скачивания из Оксфордского архива текстов как ресурс под номером 2539 [2]. В этом виде он содержит 2761 документ, каждый из которых снабжен подробной аннотацией, включающей в себя такие данные как код работы, ее название, курс, дата написания, жанр работы, учебная дисциплина, полученная оценка, количество слов. Аннотирована и информация об авторе каждой работы, в частности, такая аннотация содержит данные о поле студента, его годе рождения, первом языке, стране, откуда он родом, и т.д.

Изначально корпус создавался для исследования языковых особенностей, присущих письменным работам студентов британских высших учебных заведений [3]. В частности, по собранным в корпусе образцам изучались стиль, лексика, жанровое разнообразие академических письменных работ, зависимость стиля и жанра от области наук и дисциплины. Впоследствии корпус стал широко использоваться не только лингвистами, но и всеми, кто заинтересован в изучении письменного английского языка. Целью данной работы является демонстрация того, как корпус BAWE может использоваться в качестве лингвистического машиночитаемого ресурса при изучении курса «Введение в обработку естественного языка», преподаваемого во многих университетах мира для студентов ИТ-специальностей.

В указанной области обработки естественного языка корпус BAWE стал использоваться в качестве тестовой коллекции документов практически сразу с момента своей публикации в открытом доступе. Так, пилотная версия корпуса, состоящая всего из 500 документов, была использована в работе [4] для проведения экспериментов по автоматическому определению гендерной принадлежности авторов документов. По итогам экспериментов у 81% авторов работ пол был идентифицирован верно. В работе [5] корпус использовался для проведения экспериментов по тематическому моделированию. Авторы использовали для проверки своего метода тексты корпуса BAWE, относящиеся к области искусства и гуманитарных наук. В работе [6] авторы использовали тексты корпуса для автоматического определения тематики в английских предложениях. Разработанная этими авторами система Theme Analyzer автоматически определяла не только тема-рематическую структуру каждого предложения, но и входящие в него синтаксические узлы, тематические роли и т.д.

Одной из интересных практик применения корпуса BAWE является его использование в качестве альтернативной коллекции документов, необходимой для сопоставления с какой-либо другой коллекцией, интересующей исследователя. В работе [7] авторы используют BAWE вместе с коллекцией текстов, содержащих описания ритуальных действий, для того, чтобы извлечь ключевые слова, связанные с этой предметной областью. Авторы используют хорошо зарекомендовавший себя контрастный подход, выявляющий ключевые слова предметной области с позиции их разной встречаемости внутри предметной области и за ее пределами. Считается, что слова, которые часто употребляются внутри предметной области и крайне редко за ее пределами, являются ключевыми. В данном цитируемом случае «внутри предметной области» означает в текстах, описывающих ритуалы, а «за ее пределами» означает в текстах альтернативной коллекции, т.е. в текстах корпуса BAWE. В работе [8] авторы также используют BAWE для сравнения с другим корпусом, что по их словам, является «прямым, практичным и увлекательным способом изучения характеристик корпусов и типов текста». Авторы этой работы анализируют топ-100 ключевых слов каждого из рассматриваемых корпусов и сравнивают эти списки между собой.

Основными критериями качества альтернативной коллекции текстов являются ее представительность и сбалансированность. Представительность означает, что альтернативная коллекция по возможности должна охватывать как можно больше текстов из как можно большего числа предметных областей, не смежных с целе-

вой предметной областью. Сбалансированность означает, что различные предметные области в альтернативной коллекции по возможности должны быть представлены либо в равных пропорциях. С точки зрения названных критериев корпус BAWE является достаточно представительным (124516 словоупотреблений) и сбалансированным (4 области наук представлены примерно равными количествами текстов). В таблице 1 показано распределение текстов BAWE по областям наук, а на рисунке 1 представлена диаграмма, иллюстрирующая сбалансированность BAWE. На рисунке 2 показаны гистограммы распределения текстов корпуса в каждой из четырех областей наук с детализацией по учебным дисциплинам. Из гистограмм видно, что больше всего текстов в BAWE приходится на долю дисциплины Инженерия (238), затем идет Биология (169) и на третьем месте – Бизнес (146).

Таблица 1. Распределение текстов корпуса BAWE по областям наук

№	Область наук	Английское название	Количество текстов
1	Искусство и гуманитарные науки	Arts and Humanities (AH)	705
2	Науки о жизни	Life Sciences (LS)	683
3	Физические науки	Physical Sciences (PS)	596
4	Социальные науки	Social Sciences (SS)	777
	ИТОГО		2761

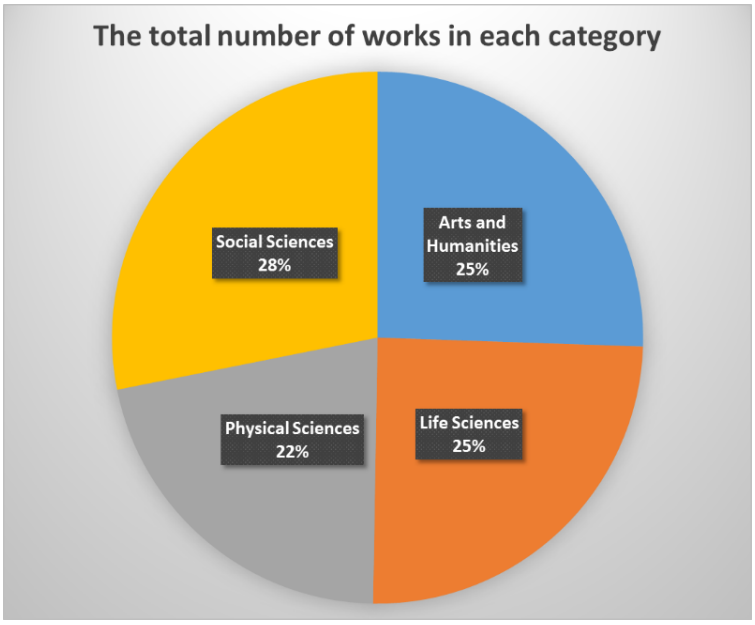

Рис. 1. Диаграмма распределения текстов корпуса BAWE по областям наук

Для семантического анализа корпуса BAWE на практических занятиях студенты использовали язык R и его библиотеки tm и quanteda [9,10]. Тексты корпуса BAWE были скачаны с сайта Оксфордского архива текстов, где они были выложена в открытом доступе в виде архива текстовых файлов [2]. Текст каждого файла был лемматизирован с помощью инструмента Wordnet Lemmatizer, входящего в состав пакета NLTK – открытой библиотеки программ для символьной и статистической об-

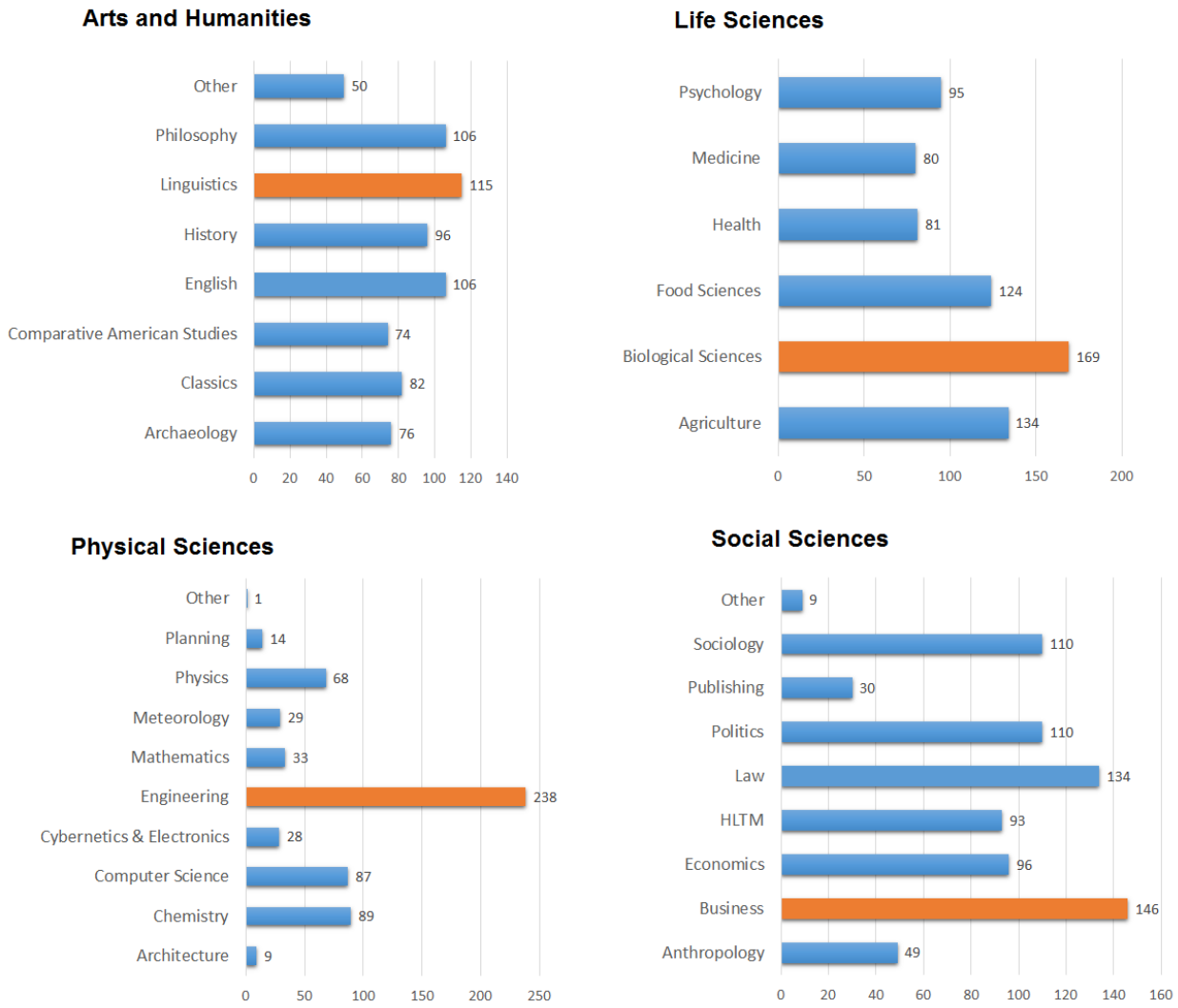

Рис. 2. Распределение текстов корпуса BAWE по учебным дисциплинам

работки естественного языка [11]. Затем эти тексты были загружены в R и преобразованы в вид удобный для обработки (представлены в виде разреженных матриц документы-на-термины).

Студентам было предложено проанализировать распределение слов в трех самых представительных дисциплинах и построить их облака (см. рисунки 3-5). Поскольку количество всех слов очень велико для визуализации, то использовались только слова с частотой употребления не меньше 70. Перед построением облаков из текстов были удалены числа, знаки пунктуации и стоп-слова. Код данной программы в R представлен ниже.

```
library(tm)
library(wordcloud)
myStoplist<-as.vector(read.csv("D:/2017/english_stopwords.csv"))
stoplist<-union(myStoplist$term, stopwords("english"))
stoplist <-c(stoplist, c("fnote", "enote", "heading"))
parent.folder <- "C:/2017/BAWE/lemmagen/PS"
sub.folders <- list.dirs(parent.folder, recursive=TRUE)[-1]
```

```
for (i in sub.folders) {
  corpus <- Corpus(DirSource(i))
  corpus<-tm_map(corpus, removeWords, stoplist)
  tdm <- TermDocumentMatrix(corpus)
  freq <- sort(rowSums(as.matrix(tdm)),decreasing=TRUE)
  set.seed(123)
  wordcloud(names(freq), freq, min.freq=70,  max.freq=Inf,
    colors=brewer.pal(8,"Dark2"))
}
```

Рис. 3. Облако слов на основе текстов дисциплины «Инженерия»

Затем студентам было предложено найти попарные пересечения топ-100 ключевых слов для каждой из трех рассматриваемых учебных дисциплин (см. таблицу 2). Попутно из числа извлеченных топ-100 слов были выделены топ-слова, общие для всех трех дисциплин (см. таблицу 3). Это такие распространенные в научных текстах слова как результат (result), система (system), фактор (factor), процесс (process), таблица (table) и т.д. Если распространить данный подход на все дисциплины корпуса BAWE, то можно говорить о перспективе автоматического построения словаря общенаучной и межотраслевой лексики на основе корпуса BAWE.

В качестве самостоятельного задания студентам было предложено найти в Интернете и изучить имеющиеся работы по автоматическому или полуавтоматическому построению словарей общенаучной лексики на основе корпусов научных текстов. В качестве примера такой работы для английского языка была указана работа [12].

Таблица 2. Пересечения по топ-словам дисциплин «Инженерия», «Биология» и «Бизнес»

Инженерия и Биология – 26 слов	Инженерия и Бизнес – 47 слов	Биология и Бизнес – 35 слов
activity, change, control, development, factor, figure, form, group, high, important, increase, level, need, number, order, process, product, production, quality, rate, result, role, study, system, table, time, year	analysis, based, business, case, change, company, control, cost, current, customer, development, factor, figure, financial, good, group, high, important, increase, information, level, management, market, model, need, number, order, performance, point, power, price, problem, process, product, profit, project, rate, result, service, strategy, system, table, team, term, time, work, year	area, change, control, data, development, effect, energy, experiment, factor, figure, formula, group, high, higher, important, increase, level, method, need, number, order, picture, process, product, rate, required, result, small, stage, system, table, temperature, time, type, year

Таблица 3. Общеакадемические топ-слова, выделенные из пересечения топ-слов дисциплин «Инженерия», «Биология» и «Бизнес»

№	Слово	№	Слово	№	Слово
1	change	8	important	15	product
2	control	9	increase	16	rate
3	development	10	level	17	result
4	factor	11	need	18	system
5	figure	12	number	19	table
6	group	13	order	20	time
7	high	14	process	21	year

На последующих занятиях корпус BAWE использовался как инструмент тематического моделирования текстов, как корпус с тестовыми данными для автоматической классификации текстов, как альтернативный корпус для извлечения ключевых слов предметной области и т.д. Таким образом, студенты смогли практически убедиться в ценности эталонных текстовых корпусов при изучении современных методов обработки естественного языка.

Литература

1. Heuboeck A. The BAWE corpus manual / A. Heuboeck, J. Holmes, H. Nesi. – Technical report, Universities of Warwick, Coventry and Reading, 2007. – 57 p.

2. Nesi H. British Academic Written English Corpus [Электронный ресурс] / H. Nesi, Sh. Gardner , P. Thompson , P. Wickens. – Режим доступа: http://ota.ahds.ac.uk/headers/2539.xml

3. Ebeling S.O. Encoding document information in a corpus of student writing: the British Academic Written English corpus / S.O. Ebeling , A. Heuboeck // Corpora. — 2007. — Vol. 2, №. 2. — P. 241–256.

4. Doyle J. Automatic categorization of author gender via n-gram analysis / J. Doyle , V. Keselj // The 6th Symposium on Natural Language Processing, SNLP. — 2005. — P. 1–5.

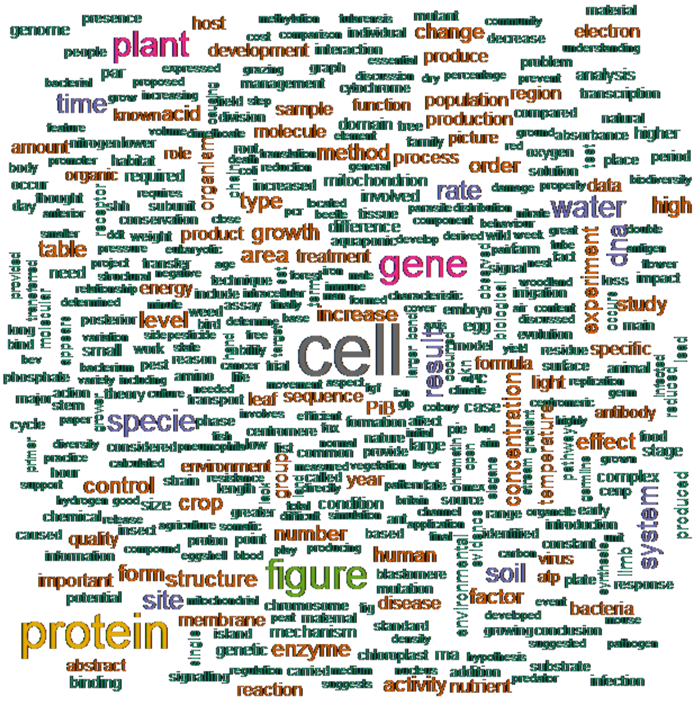

Рис. 4. Облако слов на основе текстов дисциплины «Биология»

Рис. 5. Облако слов на основе текстов дисциплины «Бизнес»

5. Allahyari M. Automatic topic labeling using ontology-based topic models / M. Allahyari , K. Kochut // Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference. — IEEE, 2015. — P. 259–264.

6. Park K. Automatic analysis of thematic structure in written English / K. Park , X. Lu //International Journal of Corpus Linguistics. — 2015. — Т. 20., №. 1. — P. 81–101.

7. Reiter N. Adapting standard NLP tools and resources to the processing of ritual descriptions / N. Reiter, O. Hellwig, A. Mishra, I. Gossmann, B. M. Larios, J. Rodrigues, B. Zeller, A. Frank // ECAI, 2010. -– 2010. -– P. 39.

8. Kilgarriff A. Getting to know your corpus / A. Kilgarriff //International Conference on Text, Speech and Dialogue. — Springer Berlin Heidelberg, 2012. — P. 3-15.

9. Feinerer I. Introduction to the tm Package Text Mining in R //2013-12-01]. http://www, dainf, ct. utfpr, edu. br/-kaestner/Min-eracao/RDataMining/tm, pdf. – 2017.

10. Benoit K. quanteda: Quantitative Analysis of Textual Data / K. Benoit , P. Nulty // An R library for managing and analyzing text. – 2013.

11. Perkins J. Python 3 Text Processing with NLTK 3 Cookbook./ J. Perkins // — Packt Publishing Ltd. – 2014. – 304 p.

12. Da Sylva L. Corpus-based derivation of a “basic scientific vocabulary” for indexing purposes / L. Da Sylva //Journal of Linguistics. – 2009. – Т. 45., №. 1. -– P. 167–201.

APPLICATION OF THE BRITISH CORPUS OF ACADEMIC WRITTEN ENGLISH IN THE COURSE «INTRODUCTION TO NATURAL LANGUAGE PROCESSING» FOR STUDENTS OF IT SPECIALTIES

!!!, E.M. Bayburin

This paper describes the using of the British Corpus of Academic Written English as a machine-readable resource illustrating the possibilities of modern methods of natural language processing. The British Corpus is a well-balanced and representative electronic collection made up of high-quality academic texts on 35 subjects.

Keywords: natural language processing, Information retrieval, corpus.

УДК 514.822

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ ВОСПРИЯТИЯ СТУДЕНТАМИ ИНФОРМАЦИИ С ПЕЧАТНОГО И ЭЛЕКТРОННОГО ИСТОЧНИКОВ

Ю.В. Пластинина[1], Т.В. Носакова[2]

1 *j.plastinina@yandex.ru*; ФГАОУ ВО «Уральский федеральный университет им. Первого Президента России Б.Н. Ельцина»

2 *nosakovatv@mail.ru*; ФГАОУ ВО «Российский государственный профессионально-педагогический университет»

В России постепенно внедряется дистанционная форма высшего образования. Она предполагает использование компьютерных технологий в процессе обучения. Однако теоретические и практические исследования восприятия студентами информации с печатного и электронного источников показали эффективность восприятия с печатных источников. При долговременном использовании компьютерными источниками в процессе образования у студентов снижаются восприятие и внимание, в также способность к абстрактному мышлению и долговременному запоминанию.

Ключевые слова: компьютерные технологии высшего образования, эффективность восприятия информации студентов, печатные и электронные источники информации.